Research in Molecular Medicine

Data Mining for Identification of Forkhead Box O3 (FOXO3a) in Different Organisms Using Nucleotide and Tandem Repeat Sequences

Sabah Mayahi[1,2,3], Ahad Yamchi[4], Masood Golalipour[3], Majid Shahbazi[3,5]*

1. Department of Molecular Medicine, School of Advanced Technologies in Medicine, Golestan University of Medical Sciences, Gorgan, Iran.
2. Department of Medical Mycology and Parasitology, School of Medicine, Mazandaran University of Medical Sciences, Sari, Iran.
3. Department of Molecular Medicine, Medical Cellular and Molecular Research Center, Golestan University of Medical Sciences, Gorgan, Iran.
4. Department of Plant Breeding and Biotechnology, Faculty of Plant Production, Gorgan University of Agricultural Sciences and Natural Resources, Gorgan, Iran.
5. Arya Tina Gene (ATG) Biopharmaceutical Company, Gorgan, Iran.

Use your device to scan and read the article online

Citation Mayahi S, Yamchi A, Golalipour M, Shahbazi M. Data Mining for Identification of Forkhead Box O3 (FOXO3a) in Different Organisms Using Nucleotide and Tandem Repeat Sequences. Research in Molecular Medicine. 2020; 8(1):17-30. https://doi.org/10.32598/rmm.8.1.17

https://doi.org/10.32598/rmm.8.1.17

Article Type:
Research Paper

Article info:
Received: 27 Nov 2019
Revised: 29 Dec 2019
Accepted: 8 Jan 2020

Keywords:
Decision tree, Data mining, Rapidminer, Forkhead box O3

ABSTRACT

Background: Deregulation of FOXO3a gene which belongs to Forkhead box O (FOXO) transcription factors, can cause cancer (e.g. breast cancer). FOXO factors have important role in ubiquitination, acetylation, de-acetylation, protein-protein interactions and phosphorylation. Understanding the regulation and mechanisms of FOXO3a can lead to cancer treatment. The aim of this study recent association of data mining with genetics has provided a strong tool for knowledge discovery.

Materials and Methods: Using genetics and bioinformatics, 30 sequences of FOXO3a genes were extracted from different species and were used in two datasets including 65 nucleotide features and 51 tandem repeat sequences. Then, we used different feature weighting and decision tree data mining algorithms on these datasets.

Results: Among nucleotide features, the frequency of AA dinucleotide was the most important genomic feature for FOXO3a gene identification. Among tandem repeat sequences, the strings of TTTTTTTTT, GAGGAGGAG, CGGCGGCGGCGG and CGGCGGCGGCGGCGG were the most effective ones to distinguish FOXO3A gene between different species.

Conclusion: The results of this study are important in understanding FOXO3a gene and developing a pathway for cancer and gene therapies in humans.

Introduction

Forkhead transcription Factors (FOXs) show specific patterns in cell cycle control, anxiety response, differentiation, and apoptosis [1]. There are more than 100 proteins in the forkhead transcription factors of the O subgroup (FOXO). This subgroup contains four members: FOXO6, FOXO4, FOXO1, and FOXO3a. FOXO proteins were first detected in certain tumors of humans at chromosomal rearrangements [2]. FOXO genes are involved in many signaling

*** Corresponding Author:**
Majid Shahbazi, PhD.
Address: *Cellular and Molecular Research Center, Golestan University of Medical Sciences, Gorgan, Iran.*
Phone: *+98 (12) 5140251*
E-mail: *shahbazimajid@yahoo.co.uk*

pathways and play crucial role in pathological and physiological processes.

Adjustment of the transcriptional and subcellular localization activity of FOXO protein is made primarily by posttranslational modifications, such as acetylation and phosphorylation [3]. The PI3K-Akt pathway is mainly involved in cellular processes including tumor suppression, cell cycle arrest, cell death, cell differentiation, metabolism, and protection against stress. Three sites on FOXO proteins are phosphorylated by Akt, resulting in nuclear exclusion and inactivation, which are related to cancer progression and tumorigenesis [4].

FKHRL1 (FOXO3) belongs to the FOX family [5]. Bioinformatics techniques (such as data mining in molecular biology, extraction of beneficial outcome from large amounts of raw data, genetics and genomics) help in sequencing and showing their observed mutations. Some of the tandem repeats such as (CT)n and (CA)n near specific genes may affect the expression of genes. Tandem repeats are used at different levels of biological structure. Most of them remain unknown; thus, for finding out their biological functions, they should be examined further. In different tumor tissues, abnormal expression of FOXO3a has been detected [1]. In this regard, we evaluated different tandem repeat sequences and nucleotide features in this study to identify FOXO3a gene in different organisms.

Materials & Methods

Database preparation and gene features

Thirty FOXO3a genes from different samples (human, animal) were extracted from the NCBI database including 65 gene features (e.g. weight/length/frequency of nucleotides, salt concentration). The information was extracted using trial CLC Main Workbench software and imported to RapidMiner GmbH 7.1.1 software.

Feature weighting

Feature weighting is a data pre-processing method and an alternative to keeping or eliminating an attribute in data mining methods, such as classification and clustering algorithms [6]. These models include 10 different operators.

Weight by information gain

The weight determines the feature's relevance to the data gain ratio and, hence, specifies the weight of a feature.

Weight by rule

Through creating a rule for every attribute and measuring the errors, the weight determines the attribute's relevance to an example set; however, this operator can be used only with a nominal label.

Weight by the average value

This operator applies a database of examples for characterizing a class by assigning weights to the features. The characteristic features are assigned higher weights in comparison with less characteristic features. For assigning a weight to a feature, the average weight value is measured for all target class examples.

Weight by deviation

In this operator, the relevance of attributes to an example set is determined based on their standard deviations. The higher weight of an attribute is an indicative of more relevance. We can normalize the standard deviations by average, maximum, or minimum of the attribute; it should be noted that this operator is applied only on ExampleSets with a numerical label.

Weight by correlation

The relevance of attributes is determined by examining the correlation for every feature of the input ExampleSet relative to the label attribute. This scheme which is based on the correlation, presents the squared or absolute value of correlation as attribute weight. This operator can be applied on sets with binominal or numerical labels.

Weight by chi-squared statistic

This operator allocates user-defined weights to the attributes, which are selected with regular expressions.

Weight by Gini index

As a measure of impurity in an ExampleSet, this scheme determines the attribute's relevance with respect to the Gini impurity index. The higher weight of an attribute is associated with greater relevance; this measure can be used for ExampleSets with nominal labels.

Weight by relief

This operator measures the attribute's relevance by relief. The main idea is to determine the attribute's quality considering how well its value discriminates between similar and different classes near each other. The rela-

tion of features is determined by sampling examples and comparing the value of the feature with the nearest examples of different and same classes. It also can be applied to various classes and regression datasets. The obtained weights are normalized into the interval between 0 and 1 if the normalized weight parameter is set to true.

Weight by support vector machine

In this operator, the attribute's relevance is determined by calculating the weight for each attribute from the input ExampleSet relative to the class attribute. The hyperplane coefficients are calculated by an Support Vector Machine (SVM) as attribute weights; this operator can be applied to multiple classes, as well.

Weight by principal component analysis

In Principal Component Analysis (PCA), an orthogonal transformation is used for converting the observations of possibly correlated attributes into the values of uncorrelated attributes called "principal components". This operator presents the weights of the attributes from the ExampleSet using a PCA-developed component and performs in the same way as a PCA model given to the weight by a component model.

Feature selection for different feature weighting operators

After measuring weighting models on each data set, each gene received a value between 0 and 1, and then the variables with higher ranks were selected.

Types of decision trees

Decision tree

A Decision Tree (DT) has a structure similar to a flowchart, where every internal node indicates a test of an attribute, and every leaf node presents a class label. The classification rules are determined by the root-to-leaf paths.

Random forest

Random decision forests [7] are an ensemble method for regression, classification, and other tasks via creating various DTs at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. These random forests correct the overfitting of DTs on their training set.

Decision stump

The tree shown in Figure 1 is a one-level DT [8]. This model has a root connected immediately to the end nodes. Prediction by a decision stump is based on the value of a single input attribute. Sometimes, these models are called 1-rules [9].

Tree induction models

The four trees including Decision Stump (DS), DT, Random Tree (RT), and Random Forest (RF) were run on 11 datasets. Each tree considers 4 criteria: information gain, gain ratio, Gini index, and accuracy. Several combinational machine learning models were used, including DT Gain Ratio, DT Information Gain, DT Gini Index, DT Accuracy, RT Information Gain, RT Gain Ratio, RT Gini Index, RT Accuracy, DS Gain Ratio, DS Information Gain, DS Gini Index, DS Accuracy, RF Gain Ratio, RF Information Gain, RF Gini Index, and RF Accuracy. In the RF models for each criterion, 10 different trees were generated. With tree induction models, 572 trees were induced.

Tandem repeat sequences

The tandem repeat sequences of nucleotides were obtained by the Microsatellite repeats finder tool. The features for all extracted gene sequences were given.

Feature selection and different attribute weighting algorithms

After running different operators of feature weighing method on the data set, each feature of the gene gained a value between 0 and 1 with respect to the target gene. We selected the variables with a weight higher than 0.5 as the best feature in the weighing model.

Creating new dataset and feature weighting algorithms for trimming the main dataset

For the gene feature and tandem repeat datasets, 10 new datasets were created containing the attributes that were important in feature weighting algorithms. These feature weighting models were: Information Gain, Gini Index, Rule, Deviation, Correlation, Chi-Squared, Uncertainty, Relief, SVM, and PCA. These models were used as predictive trees. For the induction model of trees, we used 11 datasets (10 weighing models and one main dataset).

Tree induction models and analysis

We ran all datasets with four tree induction models. As a result, 16 machine learning models were obtained mentioned in Section 2.5. For every criterion of the Random Forest model, 10 trees were created. Using tree induction models, a maximum of 572 trees was induced. For analysis of the datasets, we selected 10 important feature weighting models (Table 1).

Data analysis

The data were analyzed in GraphPad Prism v.6.0 software using Analysis of Variance (ANOVA). A p-value less than 0.05 was considered statistically significant.

Results

Dataset

The gene features dataset includes the nucleotide features of FOXO3a in different species, The dataset consists of 30 species with 65 features and 51 tandem repeat datasets. The species included Human and Xenopus (n=3 for each); Monkey, Mouse, Orangutan, Pig, Anas, and Fish (n=2 for each); Anser, Bos taurus, Dog, Gorilla, Rabbit and Ovis (n=1 for each).

Feature weighting

The data were normalized prior to the model running. Different feature weighting models were run on the datasets of tandem repeat and gene feature. The value of each features was between 0 and 1, and those with a weight greater than 0.5 were selected. Table 1 shows the important genome features in the different species. The frequency of AA dinucleotide was the most important feature which was selected by 67% of feature weighting models (Table 1). Ten sequences of tandem repeats were selected as the most important attributes which included TTTTTTTTT, GAGGAGGAG, CGGCGGCGGCGG, CGGCGGCGGCGGCGG, AAAAAAAAA, AAAAAAAAAAAA, TTTTGTTTTGTTTTG, CGGCGGCGG, AAGAAGAAG, and GGAGGAGGAGGA (Table 2).

Identification of genome feature dataset and tree induction

From 572 induced DTs with different induction models, the DT random forest was the best model when running with either information gain or Gini index criteria filtered by "chi-squared" and "correlation" feature weighting models in recognition of FOXO3a genes. Figure 1 shows the DT random forest induction model ran with the information gain criterion when running on genome

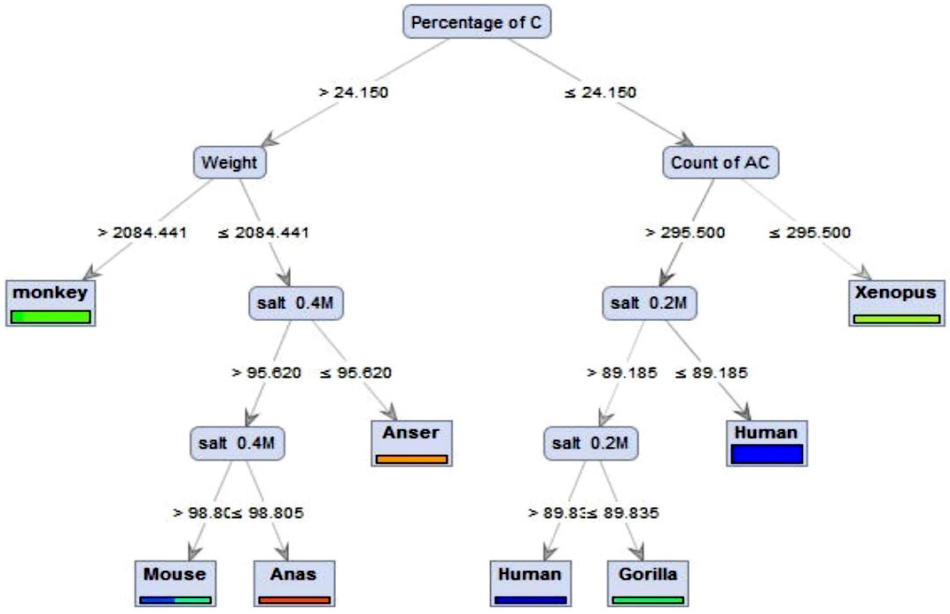

Figure 1. Identification of FOXO3a genes between different species using DT random forest induction model ran with information gain criterion and filtered by chi-squared feature weighting model

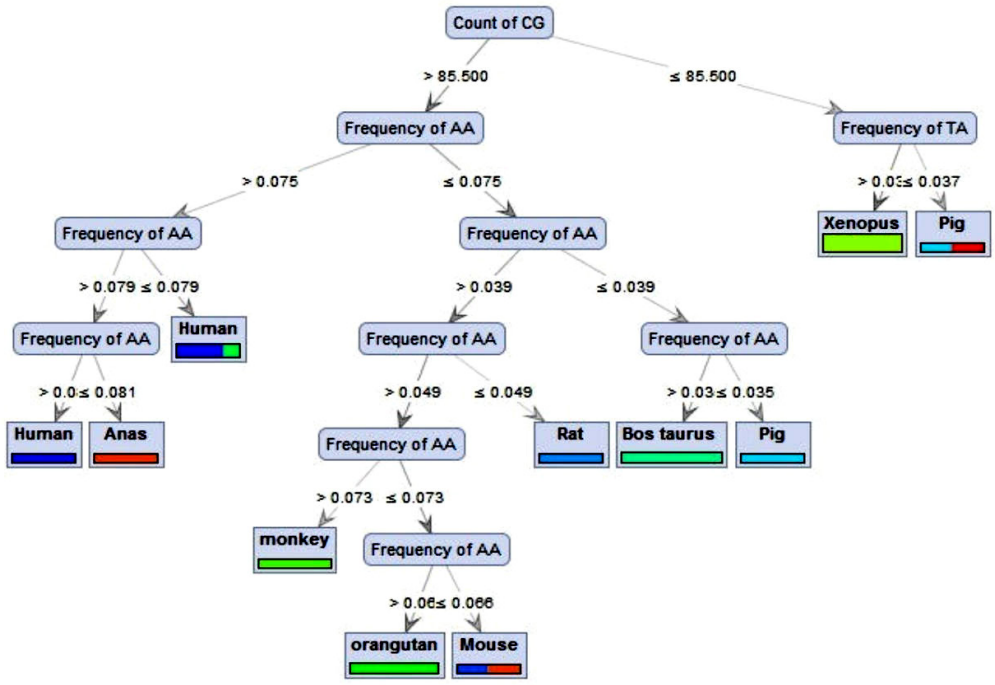

Figure 2. Identification of FOXO3a genes between different species using DT random forest induction model ran with the Gini index criterion and filtered by correlation feature weighting model

feature dataset filtered by the "chi-squared" model. This model shows the percentage of C as the main feature in FOXO3a separation. If the frequency of this feature is less than or equal to 24.150, it depends on the frequency of next feature (i.e. Count of AC). If the Count of AC is less than or equal to 295.500, the record is related to the Xenopus, but If it was more than 295.500, it belongs to salt 0.2 M. If salt 0.2 M was less than or equal to 89.185, the record belongs to human, and so on.

Figure 2 shows the DT random forest model ran with the Gini index criterion on the dataset pre-filtered by the "correlation" feature weighting model. In this model, the frequency of TA and AA dinucleotides are the important features in different sequences of FOXO3a gene. The comparison of the accuracy of the different induced tree models with feature weighting models is shown in Table 2. As can be seen, the DT models of "RF Gain Ratio" and "RF Gini Index" had the highest mean accuracy of 77.97% and 79%, respectively. Among the

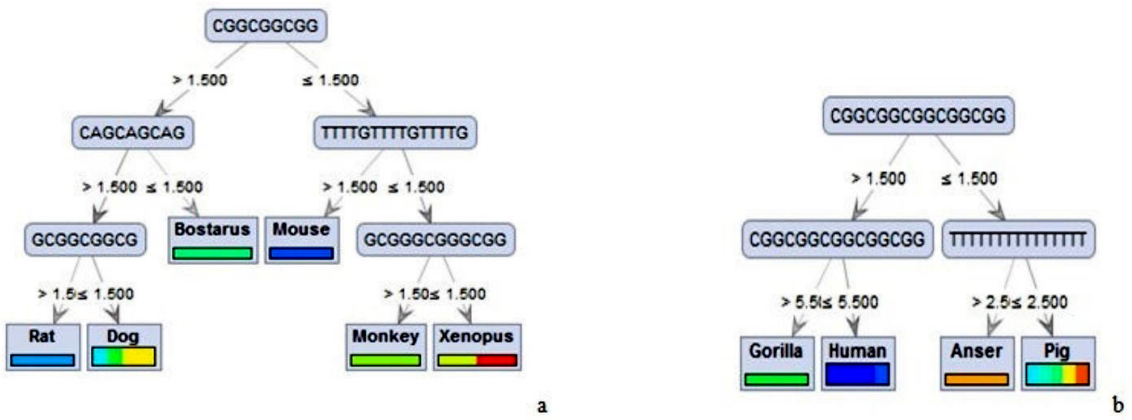

Figure 3. Identification of FOXO3a genes between different tandem repeats using DT random forest model ran with the accuracy criterion filtered by: a. rule and b. Gini index feature weighting model

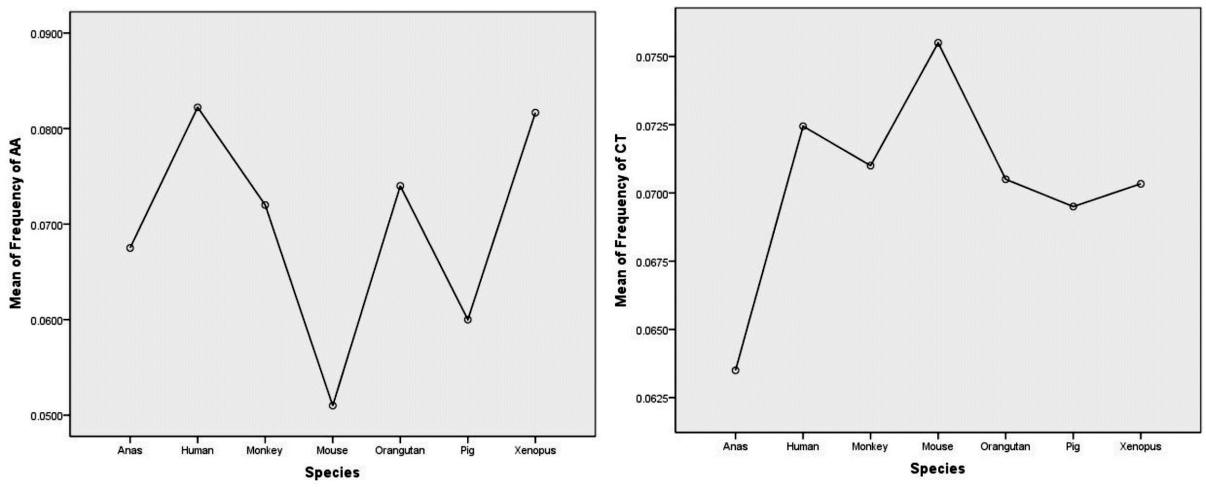

Figure 4. Plot of the mean frequency of AA and CT dinucleotide in different species

Table 1. Important genomic features of FoxO3a gene nucleotides with weights >0.5

Feature Rank	Genomic Feature	Number of Weighting Models With Important Features (Weight >0.5)
1	Frequency of AA	6
2	Frequency of CG	5
3	Frequency of CT	5
4	Frequency of CC	5
5	Frequency of TT	5
6	salt 0.2 M	5
7	salt 0.3 M	5
8	salt 0.5 M	5
9	Frequency of C + G	5
10	Frequency of A + T	5
11	Weight	5
12	Frequency of carbon	4
13	Frequency of GT	4
14	Frequency of TG	4
15	Frequency of TC	4
16	Frequency of Thymine	4
17	salt0 .1 M	4
18	salt 0.4 M	4
19	Frequency of Cytosine	4
20	Frequency of nitrogen	4
21	Frequency of AG	4
22	Frequency of oxygen	4
23	Frequency of GC	4
24	Frequency of AT	4
25	Frequency of TA	4
26	Frequency of hydrogen	4
27	Frequency of GA	4
28	Frequency of Adenine	3
29	Frequency of Guanine	3
30	Frequency of AC	3

Table 2. Important genomic features of FOXO3a tandem repeat sequences with weights >0.5

Feature Rank	Tandem Repeat Sequence	Number of Models With Important Features (Weight >0.5)
1	TTTTTTTTT	7
2	GAGGAGGAG	7
3	CGGCGGCGGCGG	7
4	CGGCGGCGGCGGCGG	7
5	AAAAAAAAA	6
6	AAAAAAAAAAAA	6
7	TTTTGTTTTGTTTTG	6
8	CGGCGGCGG	5
9	AAGAAGAAG	5
10	GGAGGAGGAGGA	5

feature weighting models for genome features, the "Correlation" and "Information Gain" models had the highest mean accuracy (60 and 59.56%, respectively). According to this table, the highest accuracy of genomic features (83.33%) was obtained by both induction models of "RF Information Gain" filtered by the chi-squared feature weighting model, and "RF Gini Index" filtered by correlation feature weighting model in comparison with the original dataset (without feature selection) and the dataset with feature selection (by feature weighting models).

Induced tree and Tandem repeat sequence features

Figure 3 shows RF tree for tandem repeat sequences when run with the accuracy criterion on a dataset filtered by the "Rule" (Fig.3a) and the "Gini Index" (Fig.3b) feature weighting models. As shown in Table 3, the FOXO3A gene sequences of TTTTTTTTT, GAGGAGGAG, CGGCGGCGGCGG, and CGGCGGCGGCGGCGG were important features in different organisms. The accuracy of tree models in recognition of FOXO3A in difference species were compared with those of different feature weighting based on the tandem repeat sequences (Table 4). The RF Information Gain and RF Accuracy had the highest mean accuracy (59.28% and 60.71%, respectively).

Original dataset vs. the dataset with feature selection

The most important point in both original (without feature selection) and the genomic dataset with feature weighting is their accuracy level. As shown in Table 3, the mean accuracy of the dataset with feature selection (by the correlation feature weighting model) increased by 1.5% compared to the dataset without feature selection (60% vs. 58.54%). For the tandem repeat dataset, Table 4 shows that the mean accuracy of the dataset with feature selection (by uncertainty attribute weighting model) increased by 4.5% compared to the dataset without feature selection (34.82% vs. 30.36%). This indicates the importance of feature weighting in prediction accuracy.

Comparative analysis of ten key features

By feature weighting, it was shown that the frequencies of TT, CG, and AA dinucleotides were highly variable in pig, anas, xenopus, and mouse species. The statistics of 10 selected main features in different organisms are shown in Table 5. This table presents the Mean±SD and Coefficient of Variance (CV). As can be seen, the frequencies of CT, A+T, C+G and CC were more variable in pig samples. These genomic key features, selected from Table 1, were subjected to ANOVA between different species. Its results showed that all selected features were significantly different among species ($P \leq 0.05$). As an example shown in Figure 4, the least mean frequency

Table 3. Comparing the accuracy of different induction trees with different feature weighting models in identification of FOXO3a in difference spices (nucleotide feature database)

Feature Weighing Models	Induction Trees (%)								
	DT Gain Ratio	DT Information Gain	DT Gini Index	DT Accuracy	RT Gain Ratio	RT Information Gain	RT Gini Index	RT Accuracy	DS Gain Ratio
Information Gain	60.00	80.00	70.00	56.67	60.00	63.33	63.33	36.67	40.00
Gini Index	60.00	80.00	73.30	56.67	46.67	63.33	60.00	36.67	40.00
Rule	56.67	80.00	73.33	50.00	60.00	70.00	70.00	36.67	36.67
Deviation	63.33	60.00	63.33	30.00	63.33	60.00	63.33	36.67	33.33
Correlation	60.00	80.00	70.00	63.33	63.33	66.67	60.00	36.67	40.00
Chi-squared statistic	60.00	80.00	70.00	56.67	43.33	63.33	70.00	36.67	40.00
Uncertainty	60.00	80.00	70.00	56.67	53.33	66.67	70.00	36.67	40.00
Relief	63.30	63.30	70.00	70.00	63.33	63.33	70.00	36.67	40.00
PCA	63.33	63.33	70.00	56.67	63.33	63.33	70.00	36.67	40.00
Wthout feature selection (original dataset)	60.00	80.00	70.00	56.67	53.33	66.67	63.33	36.67	40.00
Maximum accuracy	63.33	80.00	73.33	70.00	63.33	70.00	70.00	36.67	40.00
Minimum accuracy	56.57	63.00	63.33	30.00	43.33	60.00	60.00	36.67	33.33
Mean accuracy	60.66	74.66	70.00	55.34	57.00	64.67	66.00	36.67	39.00

Feature Weighing Models	Induction Trees (%)									
	DS Information Gain	DS Gini Index	DS Accuracy	RF Gain Ratio	RF Information Gain	RF Gini Index	RF Accuracy	Maximum Accuracy	Minimum Accuracy	Mean Accuracy
Information Gain	40.00	40.00	36.67	70.00	83.00	83.30	70.00	83.30	36.67	59.56
Gini Index	40.00	40.00	36.67	63.33	80.00	80.00	70.00	80.00	36.67	57.92
Rule	40.00	36.67	36.67	63.33	76.67	73.33	63.33	80.00	36.67	57.71
Deviation	36.67	36.67	36.67	60.00	70.00	76.67	60.00	76.67	33.33	53.13
Correlation	40.00	40.00	36.67	73.33	83.33	83.33	63.33	83.33	36.67	60.00
Chi-squared statistic	40.00	40.00	36.67	70.00	83.33	76.67	66.67	83.33	36.67	58.33
Uncertainty	40.00	40.00	36.67	73.33	80.00	80.00	63.33	80.00	36.67	59.17
Relief	40.00	40.00	33.33	56.67	76.67	80.00	83.33	83.33	33.33	59.37
PCA	40.00	40.00	33.33	70.00	73.33	76.67	73.33	76.67	33.30	58.33
Without feature selection (original dataset)	40.00	40.00	36.67	70.00	73.33	80.00	70.00	80.00	36.67	58.54
Maximum accuracy	40.00	40.00	36.67	73.33	83.33	83.33	83.33			
Minimum accuracy	36.67	36.67	33.33	56.67	70.00	73.33	60.00			
Mean accuracy	39.67	39.33	36.00	67.00	77.97	79.00	68.33			

Table 4. Comparing the accuracy of different induction trees with different feature weighting models in identification of FOXO3a in difference spices (Tandem repeat sequences)

Feature Weighting Models	Induction Trees (%)								
	DT Gain Ratio	DT Information Gain	DT Gini Index	DT Accuracy	RT Gain Ratio	RT Information Gain	RT Gini Index	RT Accuracy	DS Gain Ratio
Information Gain	7.04	42.86	7.14	7.14	14.29	28.57	21.43	14.29	14.29
Gini Index	35.71	42.86	7.14	7.14	14.29	21.43	21.43	7.14	14.29
Rule	35.71	42.86	7.14	7.14	21.43	28.57	28.57	7.14	14.29
Deviation	35.71	35.71	7.14	7.14	21.43	21.43	21.43	14.29	14.29
Correlation	42.86	42.86	7.14	7.14	14.29	14.29	14.29	7.14	14.29
Chi-squre static	21.43	28.57	7.14	7.14	21.43	28.57	28.57	21.43	14.29
Uncertainty	35.71	42.86	7.14	7.14	28.57	28.57	28.57	28.57	14.29
Relief	28.57	28.57	7.14	7.14	21.43	21.43	21.43	21.43	14.29
PCA	7.14	21.43	7.14	7.14	7.14	21.43	21.43	7.14	14.29
SVM	35.71	28.57	7.14	7.14	21.43	21.43	21.43	21.43	14.29
Without feature selection (original dataset)	35.71	42.86	7.14	7.14	21.43	28.57	28.57	7.14	14.29
Maximum accuracy	35.71	42.86	7.14	7.14	21.43	28.57	28.57	28.57	14.29
Minimum accuracy	7.14	21.43	7.14	7.14	7.14	14.29	14.29	7.14	14.29
Mean accuracy	28.56	35.72	7.14	7.14	18.57	23.57	22.86	15.00	14.29

Feature Weighting Models	Induction Trees (%)									
	DS Information Gain	DS Gini Index	DS Accuracy	RF Gain Ratio	RF Information Gain	RF Gini Index	RF Accuracy	Maximum Accuracy	Minimum Accuracy	Mean Accuracy
Information Gain	14.29	14.29	14.29	57.14	78.57	71.43	71.43	78.57	7.04	29.91
Gini Index	14.29	14.29	14.29	50.00	57.14	64.29	78.57	78.57	7.14	29.02
Rule	14.29	14.29	14.29	64.29	71.43	64.29	78.57	78.57	7.14	32.14
Deviation	14.29	14.29	14.29	50.00	35.71	42.86	42.86	50.00	7.14	24.55
Correlation	14.29	14.29	14.29	50.00	57.14	57.14	57.14	57.14	7.14	26.79
Chi-squre static	14.29	14.29	14.29	35.71	35.71	35.71	35.71	35.71	7.14	22.77
Uncertainty	14.29	14.29	14.29	64.29	78.57	71.43	78.57	78.57	7.14	34.82
Relief	14.29	14.29	14.29	50.00	57.14	57.14	57.14	57.14	14.29	27.23
PCA	14.29	14.29	14.29	42.86	42.86	42.86	42.86	42.86	7.14	20.54
SVM	14.29	14.29	14.29	64.29	78.57	71.43	64.29	78.57	14.29	31.25
Without feature selection (original dataset)	14.29	14.29	14.29	57.14	57.14	57.14	78.57	78.57	7.14	30.36
Maximum accuracy	14.29	14.29	14.29	64.29	78.57	71.43	78.57			
Minimum accuracy	14.29	14.29	14.29	35.71	35.71	35.71	35.71			
Mean accuracy	14.29	14.29	14.29	52.86	59.28	57.86	60.71			

Table 5. Statistics of 10 main genomic features for different species selected by feature weighting models

Genomic Feature	Species	Mean±SD	Variance	CV
Frequency of AA	Anas	0.0675±0.0177	0.0003	26.22
	Human	0.0822±0.0032	0.0000	3.89
	Monkey	0.072±0.0028	0.0000	3.89
	Mouse	0.051±0.0113	0.0001	22.16
	Orangutan	0.074±0.0014	0.0000	1.89
	Pig	0.06±0.0368	0.0014	61.33
	Xenopus	0.0817±0.0059	0.0000	7.22
Frequency of CG	Anas	0.05±0.0057	0.0000	11.40
	Human	0.0203±0.0068	0.0000	33.50
	Monkey	0.034±0.0042	0.0000	12.35
	Mouse	0.0515±0.0163	0.0003	31.65
	Orangutan	0.0345±0.0035	0.0000	10.14
	Pig	0.055±0.0283	0.0008	51.45
	Xenopus	0.0187±0.0067	0.0000	35.83
Frequency of CT	Anas	0.0635±0.0007	0.0000	1.10
	Human	0.0724±0.0022	0.0000	3.04
	Monkey	0.071±0.0000	0.0000	0.00
	Mouse	0.0755±0.0007	0.0000	0.93
	Orangutan	0.0705±0.0007	0.0000	0.99
	Pig	0.0695±0.0092	0.0001	13.24
	Xenopus	0.0703±0.0006	0.0000	0.85
Frequency of A + T	Anas	0.462±0.0283	0.0008	6.13
	Human	0.5381±0.0167	0.0003	3.10
	Monkey	0.5015±0.012	0.0001	2.39
	Mouse	0.4295±0.0544	0.0030	12.67
	Orangutan	0.5005±0.012	0.0001	2.40
	Pig	0.425±0.1004	0.0101	23.62
	Xenopus	0.5557±0.034	0.0012	6.12
Frequency of TT	Anas	0.055±0.0127	0.0002	23.09
	Human	0.0848±0.0058	0.0000	6.84
	Monkey	0.077±0.0028	0.0000	3.64
	Mouse	0.0495±0.0191	0.0004	38.59
	Orangutan	0.0755±0.0021	0.0000	2.78
	Pig	0.051±0.041	0.0017	80.39
	Xenopus	0.0860±0.0125	0.0002	14.53

Genomic Feature	Species	Mean±SD	Variance	CV
Salt 0.2 M	Anas	91.945±1.1667	1.3612	1.27
	Human	88.8289±0.683	0.4665	0.77
	Monkey	90.33±0.495	0.2450	0.55
	Mouse	93.29±2.2062	4.8673	2.36
	Orangutan	90.37±0.495	0.2450	0.55
	Pig	93.465±4.1224	16.9942	4.41
	Xenopus	88.1033±1.3967	1.9508	1.59
Salt 0.3 M	Anas	94.865±1.1667	1.3612	1.23
	Human	91.7522±0.6839	0.4677	0.75
	Monkey	93.255±0.502	0.2520	0.54
	Mouse	96.21±2.2062	4.8673	2.29
	Orangutan	93.295±0.4879	0.2380	0.52
	Pig	96.385±4.1224	16.9942	4.28
	Xenopus	91.0267±1.3929	1.9402	1.53
Salt 0.5 M	Anas	98.55±1.1597	1.3449	1.18
	Human	95.4333±0.6837	0.4674	0.72
	Monkey	96.935±0.502	0.252	0.52
	Mouse	99.89±2.2062	4.8673	2.21
	Orangutan	96.975±0.4879	0.238	0.50
	Pig	100.065±4.1224	16.9942	4.12
	Xenopus	94.71±1.3986	1.9561	1.48
Frequency of C + G	Anas	0.538±0.0283	0.0008	5.26
	Human	0.4619±0.0167	0.0003	3.62
	Monkey	0.4985±0.012	0.0001	2.41
	Mouse	0.5705±0.0544	0.003	9.54
	Orangutan	0.4995±0.012	0.0001	2.40
	Pig	0.575±0.1004	0.0101	17.46
	Xenopus	0.4443±0.034	0.0012	7.65
Frequency of CC	Anas	0.0795±0.012	0.0001	15.09
	Human	0.0638±0.0036	0.0000	5.64
	Monkey	0.0725±0.0035	0.0000	4.83
	Mouse	0.0915±0.0191	0.0004	20.87
	Orangutan	0.0735±0.0049	0.0000	6.67
	Pig	0.0895±0.0205	0.0004	22.91
	Xenopus	0.055±0.0053	0.0000	9.64

of AA and the highest mean frequency of CT dinucleotide were observed in mouse species.

Discussion

FOXO3a gene belongs to the O subclass of FOXs. These factors have different roles in a wide range of physiological processes such as tumor suppression, cellular differentiation, cell cycle arrest, metabolism, protection against stress, and cell death [10]. The deregulation of FOXO3a is associated with tumorigenesis. Its activity is often seen in cancers. It is a valuable target for cancer and gene therapies, suggesting that therapies might be effective in blocking tumor expansion and metastasis [11]. Genetics and genomics help in sequencing and finding their mutations.

KayvanJoo et al. (2014) showed that bioinformatics and nucleotide attributes of hepatitis C virus (e.g. count of hydrogen and CG) is associated with treatment outcome [12]. Tahrokh et al. (2011) by study of a large number of structural protein specification, showed that data mining algorithms is a novel functional strategy for studying the evolution [13]. Two different databases were used in our study to find a method for examining the structural differences in the Foxo3a gene of different organism; one database was based on nucleotide features and one based on the tandem repeat sequences of the gene. For this study, we used feature weighting algorithms.

These algorithms showed the importance of each feature in different organisms. In these algorithms, it was shown that the sequence of CGGCGGCGGCGGCGG is an important feature to build the trees to distinguish between gorilla, human and other organisms. The dinucleotide frequency is important in the phylogenetic structure of Foxo3a genes. The FOXO3a gene has important roles in different organisms. Recognition of FOXO3a gene is a critical step for identification. expression and regulation. Identifying the expression of human FOXO3a gene can provide information on the spread of cancer cells. Therefore, it can be identified with new criteria based on bioinformatics and genomic properties in different organisms that are very important for therapeutic purposes, such as cancer and gene therapies.

Ethical Considerations

Compliance with ethical guidelines

All ethical principles were considered in this article. The participants were informed about the purpose of the research and its implementation stages; they were also assured about the confidentiality of their information; Moreover, They were allowed to leave the study whenever they wish, and if desired, the results of the research would be available to them.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Authors contribution's

All authors contributed equally in preparing all parts of the research.

Conflict of interest

The authors declared no conflict of interest.

Acknowledgements

The authors would like to thank the Golestan University of Medical Sciences and Arya Tina Gene Biopharmaceutical Company in Gorgan, Iran for their valuable cooperation.

References

[1] Kong W, He L, Coppola M, Guo J, Esposito NN, Coppola D, Cheng JQ. MicroRNA-155 regulates cell survival, growth, and chemosensitivity by targeting FOXO3a in breast cancer. J Biol Chem. 2010; 285(23):17869-79. [DOI:10.1074/jbc.M110.101055] [PMID] [PMCID]

[2] Obsil T, Obsilova V. Structure/function relationships underlying regulation of FOXO transcription factors. Oncogene. 2008; 27(16):2263-75. [DOI:10.1038/onc.2008.20] [PMID]

[3] Fu Z, Tindall DJ. FOXOs, cancer and regulation of apoptosis. Oncogene. 2008; 27(16):2312-9. [DOI:10.1038/onc.2008.24] [PMID] [PMCID]

[4] Zhang X, Rielland M, Yalcin S, Ghaffari S. Regulation and function of foxo transcription factors in normal and cancer stem cells: What have we learned? Curr Drug Targets. 2011; 12(9):1267-83. [DOI:10.2174/138945011796150325] [PMID]

[5] Chen J1, Gomes AR, Monteiro LJ, Wong SY, Wu LH, Ng TT. Constitutively Nuclear FOXO3a Localization Predicts Poor Survival and Promotes Akt Phosphorylation in Breast Cancer. PLOS One 2010; 5(8):e12293. [DOI:10.1371/journal.pone.0012293] [PMID] [PMCID]

[6] Song YC, Meng HD, O'grady MJ, O'Hare G. Applications of attributes weighting in data mining. InIEEE Proc. of SMC UK &RI 6th Conference on Cybernetic Systems. 2007; 2007:41-5.

[7] Tin Kam Ho. Random decision forests. Proceedings of 3rd International Conference on Document Analysis and Recognition, Montreal, Quebec, Canada, 1995; 1995:278-82. [DOI:10.1109/ICDAR.1995.598994]

[8] Wayne I, Langley P. Induction of one-level decision trees. Machine learning proceedings. Edinburgh: Elsevier; 1992. [DOI:10.1016/B978-1-55860-247-2.50035-8]

[9] Holte RC. Very simple classification rules perform well on most commonly used datasets. Mach Learn. 1993; 11(1):63-90. [DOI:10.1023/A:1022631118932]

[10] Arden KC. FOXO animal models reveal a variety of diverse roles for FOXO transcription factors. Oncogene. 2008; 27(16):2345-50. [DOI:10.1038/onc.2008.27] [PMID]

[11] Storz P, Döppler H, Copland JA, Simpson KJ, Toker A. FOXO3a promotes tumor cell invasion through the induction of matrix metalloproteinases. Mol Cell Biol. 2009; 29(18):4906-17. [DOI:10.1128/MCB.00077-09] [PMID] [PMCID]

[12] KayvanJoo AH, Ebrahimi M, Haqshenas G. Prediction of hepatitis C virus interferon/ribavirin therapy outcome based on viral nucleotide attributes using machine learning algorithms. BMC Res Notes. 2014; 7:565. [DOI:10.1186/1756-0500-7-565] [PMID] [PMCID]

[13] Tahrokh E, Ebrahimi M, Ebrahimi M, Zamansani F, Sarvestani NR, Mohammadi-Dehcheshmeh M, et al. Comparative study of ammonium transporters in different organisms by study of a large number of structural protein features via data mining algorithms. Genes & Genomics. 2011; 33:565-75. [DOI:10.1007/s13258-011-0057-6]

This Page Intentionally Left Blank